AD-A095 025 TEXAS UNIV AT AUSTIN CENTER FOR CYBERNETIC STUDIES F/G 12/1
UNCONSTRAINED MINIMIZATION BY INTERPOLATION: RATES OF CONVERGEN--ETC(U)
DEC 80 J BARZILAI N00014-80-C-0242
UNCLASSIFIED CCS-389 NL

1 OF 1

AD A
095025

END
DATE
FILMED
3-81
DTIC

LEVEL II

12

AD A095025

81 2 17 053

12

Research Report CCS 389

UNCONSTRAINED MINIMIZATION BY INTERPOLATION: RATES OF CONVERGENCE

by

J. Barzilai

December 1980

This research was partly supported by Project NR047-021, ONR Contract N00014-80-C-0242 with the Center for Cybernetic Studies, The University of Texas at Austin. Reproduction in whole or in part is permitted for any purpose of the United States Government.

CENTER FOR CYBERNETIC STUDIES

A. Charnes, Director
Business-Economics Building, 203E
The University of Texas at Austin
Austin, TX 78712
(512) 471-1821

ABSTRACT

Newton's method for finding a stationary point of $f: R^n \to R$ consists of the iteration $x_{i+1} = x_i - [\nabla^2 f(x_i)]^{-1} \cdot \nabla f(x_i)$. Its main attraction is its quadratic convergence. However, it necessitates computation and inversion of the second order derivative matrix.

Common minimization algorithms approximate the Hessian or its inverse by first order (i.e., gradient) information. First order information algorithms in common use are known to have at best superlinear rate of convergence.

We analyze the rate of convergence of a new class of algorithms based on n-dimensional interpolation. In particular, we present a class of algorithms which use first order information only, while maintaining quadratic convergence.

KEY WORDS

Unconstrained minimization, Nonpolynomial interpolation, Convergence rates.

1. Introduction

Most of the commonly used algorithms for the unconstrained minimization of $f:R^n \to R$ for $n>1$, are descent methods. These are based on the iteration $x_{i+1} = x_i + \alpha_i d_i$, where $d_i \varepsilon R^n$ is a search direction, and $\alpha_i \varepsilon R$ is the step-size, usually determined by a line search as a minimizer of $f(x_i + \alpha d_i)$ over $\alpha > 0$.

The exception is Newton's method, which is based on interpolating f by a quadratic and minimizing this quadratic at each step of the algorithm.

For a general discussion of unconstrained minimization techniques see [1,8]. The following points are relevant to our discussion.

The classical steepest descent method uses first order (i.e., gradient) information only. Its main drawback is its linear rate of convergence.

Newton's method converges quadratically. However, it necessitates the costly computation and inversion of the Hessian at each iteration.

Other methods based on first order information are known to converge at best superlinearly (e.g., [6]).

Many of these methods approximate Newton's method in the sense that the search direction they generate can be shown to be a direction along which an appropriate quadratic is minimized.

A different approach is based on the unfounded assumption that algorithms having the finite termination property (i.e., solution in a finite number of steps) for a class of functions wider than the class of quadratics, are faster than those having the quadratic termination property. Thus, Jacobson and Oksman [7] generalize from quadratic termination to homogeneous functions termination. This was further generalized (see e.g., [4]).

Another step toward discarding the quadratic model has recently been taken by Davidon [5]. His motivation, however, partly coincides with ours.

In this paper, we analyze the rate of convergence of n-dimensional interpolation algorithms for unconstrained minimization. We note the following.

Most of the commonly used algorithms for one-dimensional minimization are based on polynomial interpolation. It is well known that Newton's method in this case is inefficient in the sense that quadratic convergence can be achieved using first order information only (e.g., [8, p. 142]), by using two interpolation points rather than one. This should make one doubt whether Newton's method is a suitable model for efficient algorithms.

Quadratics are inadequate for n-dimensional two-point interpolation with zero and first order information (see Davidon [5]). Therefore, non-polynomial interpolation is necessary. Our analysis shows that the rate of convergence is independent of the interpolating function. For interpolatory algorithms, therefore, the question whether the search direction coincides with Newton's direction or generalizes it, is irrelevant to the rate of convergence analysis. The same is true for termination properties.

The main difficulty in the analysis of n-dimensional interpolation algorithms is that the formulas for the error in n-dimensional interpolation are not suitable for this purpose. We overcome this difficulty by relating the n-dimensional problem to an appropriate one-dimensional interpolation problem.

2. Minimization by Interpolation

The interpolation algorithm we study generates a sequence $\{x_i\}$ as follows. Let $s \geq 1$, $m \geq 0$ be fixed integers, and let $T:R^n \to R$ depend on $r = s(m+1)$ parameters. Given m+1 approximants $x_0, \ldots, x_{n+1}$ to the solution of

(1) $\nabla f(x^*) = 0,$

we use x_i, $x_{i-1}, \ldots,$ x_{i-m} to construct a new approximant x_{i+1}. First we interpolate f by T requiring

(2) $T^{(k)}(x_{i-j}) = f^{(k)}(x_{i-j}) \qquad j=0,\ldots,m;\ k=0,\ldots,s-1.$

Here $f^{(1)} = \nabla f$, $f^{(2)} = \nabla^2 f$ etc. The new point x_{i+1} is determined by

(3) $\nabla T(x_{i+1}) = 0.$

In the following, we assume that equations (1)-(3) have solutions.

We define the rate (or order) of convergence of a sequence $\{x_i\}$ converging to x^* as the number p (if it exists) such that

$$\frac{\| x_{i+1} - x^* \|}{\| x_i - x^* \|^p} \to C \neq 0.$$

Here $\| \cdot \|$ is a fixed arbitrary norm. Ortega and Rheinboldt [9, § 9] refer to the rate p defined above as the C-order of the sequence $\{x_i\}$. When it exists, it coincides with their Q- and R- orders. We will unify our results for the C-, Q- and R- orders through the use of the C-order of convergence.

We derive the rate of convergence of the n-dimensional interpolating algorithm by establishing some difference relations for the errors $\|x_i - x^*\|$.

To derive the basic difference relation we need, we pass a curve in R^n through the points x^* and $x_{i+1}, x_i, \ldots, x_{i-m}$, i.e., we determine a function $\psi: R \to R^n$ such that

$$(4) \quad \begin{aligned} \psi(t_{i-j}) &= x_{i-j} \qquad j = -1, 0, 1, \ldots, m, \\ \psi(t^*) &= x^*, \end{aligned}$$

where the parameter t is chosen so that

$$(5) \quad t_{i-j} = \| x_{i-j} - x^* \| , \qquad t^* = \| x^* - x^* \| = 0.$$

We will later discuss this construction. Note, however, that the construction of ψ is a part of the analysis of the properties of the algorithm, not a part of the algorithm itself.

Now define $\theta(t) = T(\psi(t))$, $\phi(t) = f(\psi(t))$. Equations (2) and (4) imply

$$(6) \quad \theta^{(k)}(t_{i-j}) = \phi^{(k)}(t_{i-j}) , \qquad j = 0, \ldots, m; \; k = 0, \ldots, s-1.$$

It follows that (6) defines one-dimensional interpolation for which convenient error formula exists (see Ostrowski [10, p. 12]). Henceforth, we assume $\theta, \phi \in C^{r+1}$ in a neighborhood of $t^* = 0$. Using the one-dimensional error formula we have

$$(7) \quad \phi(t) - \theta(t) = \frac{\phi^{(r)}(\xi) - \theta^{(r)}(\xi)}{r!} \prod_{j=0}^{m} (t - t_{i-j})^s ,$$

where ψ is a point in the interval determined by $t, t_i, \ldots, t_{i-m}$. Note that formula (7) holds for general (not necessarily polynomial) interpolation.

We now differentiate (7) and set $t = 0$. From (1) and (2) we have $\phi'(o) = \theta'(t_{i+1}) = 0$, so that

$\phi'(o) - \theta'(o) = -\theta'(o) = \theta'(t_{i+1}) - \theta'(o) = t_{i+1}\, \theta''(\zeta)$, where ζ is a point between t_{i+1} and 0. Differentiating the right hand side of (7) using Ralston's result [11] on the differentiation of the error term generalized for the hyperosculatory case (see [2]), we finally have

<u>Lemma 1</u> Under the assumptions made above, the errors in the n-dimensional interpolation algorithm satisfy the difference relation

$$(8) \quad t_{i+1} = M_i \sum_{k=0}^{m} t_{i-k}^{s-1} \prod_{\substack{j=0 \\ j \neq k}}^{m} t_{i-j}^{s} + N_i \prod_{j=0}^{m} t_{i-j}^{s} ,$$

where

$$M_i = \frac{M(\xi_i(t_{i+1}))(r-1)\cdot s}{\theta''(\zeta(t_{i+1}))} , \qquad N_i = \frac{N(\eta_i(t_{i+1}))\cdot(-1)^r}{\theta''(\zeta(t_{i+1}))} ,$$

$$M(t) = \frac{\phi^{(r)}(t) - \theta^{(r)}(t)}{r!} , \qquad N(t) = \frac{\phi^{(r+1)}(t) - \theta^{(r+1)}(t)}{(r+1)!} ,$$

and where $\xi_i(t)$, $\eta_i(t)$ are in the interval determined by $t, t_{i+1}, \ldots, t_{i-m}$, and $\zeta(t_{i+1})$ is in the interval determined by $t_{i+1}, 0$. □

It follows from (8) that if the initial errors $t_0, \ldots, t_n$ are small enough, and if the coefficients $\{M_i\}$, $\{N_i\}$ are bounded, the sequence $\{t_i\}$ converges to zero, i.e., $x_i \to x^*$. Moreover, if $s \geq 2$, (8) implies

$$(9) \quad \frac{t_{i+1}}{t_i} \to 0,$$

(*i.e.*, superlinear convergence). If $s=1$, we assume $m \geq 2$. For $m=2$, (8) is the basic difference relation governing the behavior of the Quadratic Fit algorithm, which is known to converge superlinearly (see Theorem 3.4.1 in Brent [3]). It is evident from (8) that the rate of $m>2$ is not less than the rate for $m=2$. Therefore, (9) holds for all $s \geq 1$, $m \geq 0$ if $r=s(m+1) \geq 3$. Rewriting (8) in the form

$$t_{i+1} = M_i \, t_i^{s-1} \prod_{j=1}^{m} t_{i-j}^{s} \left[1 + \sum_{k=1}^{n} \frac{t_i}{t_{i-k}} + \frac{N_i}{M_i} t_i \right] ,$$

we finally have

<u>Lemma 2</u> Under the assumptions made above, and if $M_i \to M \neq 0$, the sequence t_i satisfies the difference relation

$$(10) \quad t_{i+1} = A_{i+1} \, t_i^{s-1} \prod_{j=1}^{m} t_{i-j}^{s}$$

with $A_{i+1} \to M$. □

Defining $y_i = \log|t_i|$ and $B_i = \log|A_i|$, (10) implies the difference equation

$$y_{i+1} - (s-1) \, y_i - s \sum_{j=1}^{m} y_{i-j} = B_{i+1}$$

with indicial equation

$$(11) \quad t^{m+1} - (s-1) \, t^m - s \sum_{j=0}^{m-1} t^j = 0,$$

where the sum in (11) is taken as zero if $m=0$.

Tamir [12,13] proves that under our assumptions the C-rate of convergence of the sequence $\{t_i\}$ (hence $\{x_i\}$) is given by the unique positive root of the indicial equation (11). In this case, the C-, Q-, and R-rates of convergence are exactly p, where p is the positive solution of (11).

If the limit of M_i exists and is zero, or if this limit does not exist, but the sequence $\{M_i\}$ is bounded, equation (8) implies that the Q- and R- rates of convergence are still at least p.

We summarize our results:

Theorem: Under the above assumptions, if the sequence $\{M_i\}$ is bounded, and if the initial errors of the interpolation algorithm are small enough, the sequence $\{x_i\}$ converges to the solution x^* with C- (when it exists), Q- and R- rates of convergence at least p, where p is the unique positive solution of (11).

Corollary: The rate of convergence of the sequence generated by the interpolation algorithm is independent of the interpolating function.

Remark: In order for the sequence $\{M_i\}$ to be bounded, it is sufficient that $\theta^{(r+1)}$ and $\phi^{(r+1)}$ exist and are continuous, and $\phi''(0) \neq 0$. This would be the case if f has continuous derivatives of order r+1, the parameters of T depend continuously on the data through (2), T has continuous derivatives of order r+1 for the appropriate values of the parameters, and $\phi''(0) = \dot{\psi}(0)^T \nabla^2 f(x^*) \dot{\psi}(0) \neq 0$. Setting $\psi_k(t) = \sum_{j=0}^{r} a_{jk} t^j$ $(k=1,\dots,n)$, and assuming $\frac{\partial^2 f(x^*)}{\partial x_1^2} \neq 0$, the choice $a_{11} = 1$ and $a_{ok} = 1$ for $k=2,\dots,n$ will ensure $\phi''(0) \neq 0$.

Summary and Conclusion

We have shown that the rate of convergence of n-dimensional interpolation algorithms is inherited from the underlying one-dimensional interpolation, that it is independent of the interpolating function, and is given by the unique solution of the equation

$$(12)\quad t^{m+1} - (s-1)\, t^{m} - s\sum_{j=0}^{m-1} t^{j} = 0 \quad ,$$

where m+1 interpolation points and s derivatives (of orders zero to s-1) are used.

Our work is based on the results of Traub [14] and Ostrowski [10] for the one-dimensional root-finding problem. Tamir [12,13] adapted these results for the minimization problem. In [12] he studies the rate of convergence of algorithms using function values only (m=0) with a superfluous assumption and a false conjecture. This detailed analysis is repeated in [13] for the case m>0. He treats polynomial interpolation only, and shows that for fixed s and $m\to\infty$, the rate p tends to

$$(13)\quad \frac{s}{2} + \sqrt{\left(\frac{s}{2}\right)^2 + 1.}$$

However, he neglects to realize the effect of memory on the rate of convergence, which is implied by (11) and (12).

Indeed, for fixed s, the rate is obtained for m=0 and m=1 by solving the indicial equations $t-(s-1)=0$ and $t^2-(s-1)t-s=0$, respectively. Therefore, $p=s-1$ for m=0, $p=s$ for m=1 and $p \to \frac{s}{2}+\sqrt{\left(\frac{s}{2}\right)^2+1}$ for $m\to\infty$. It follows that algorithms using more than two interpolation points are inefficient, and two-point algorithms are substantially faster than one-point algorithms.

In particular for $m=1$ and $s=2$ we have a two-point algorithm using first-order information with second-order rate of convergence (which is a well-known result in the one-dimensional case).

Note that no line search is needed in this class of algorithms, and that they may be designed to locate saddle points rather than minimum points. A line search, however, may serve as part of a globalizing procedure.

Compare also the discussion in Davidon [5] regarding the difficulty of determining the effect of memory on the performance of descent algorithms.

We have not computed the asymptotic error constant, since it depends on the norm used (see Ortega and Rheinboldt [9]). This can be computed, however, under the appropriate assumptions (cf. Tamir [12,13]). We have also made no attempt at giving the strongest results (i.e., the weakest assumptions) possible. Compare, for example, Brent [3].

Finally, we remark that the interpolation requirements (2) leave much freedom for the choice of the interpolation function T. A useful choice for it seems to be a separable sum of rational functions $T(s) = \sum_{k=1}^{n} R(x_k)$. For example, for $s=2$, $m=1$, one can choose $R(x) = \frac{ax^2+bx+c}{dx-1}$, with the values of components of T equal to one another at each interpolation point.

ACKNOWLEDGMENTS

The author is indebted to Dr. A. Charnes for his unfailing support, and to Uzi Ravhon for too many things to list.

REFERENCES

[1] Avriel, M., Nonlinear Programming, Analysis and Methods, Prentice-Hall, Englewood Cliffs, N.J., 1976.

(2] Barzilai, J., "Convergent Methods for Optimization Problems," Ph.D. Dissertation, Technion, Israel Institute of Technology, Haifa, Israel, 1980.

[3] Brent, R.P., Algorithms for Minimization without Derivatives, Prentice-Hall, Englewood Cliffs, N.J., 1973.

[4] Charalambous, C., "Unconstrained Optimization Based on Homogeneous Models," Mathematical Programming 5(1973) 189-198.

[5] Davidon, W.C., "Conic Approximations and Collinear Scaling for Optimizers," SIAM J. Numer. Anal., 17(1980) 268-281.

[6] Dennis, J.E., and Moré, J.J., "Quasi-Newton Methods, Motivation and Theory," SIAM Review 19(1977) 46-87.

[7] Jacobson, D., and Oksman, W., "An Algorithm that Minimizes Homogeneous Functions of N Variables in N+2 Iterations and Rapidly Minimizes General Functions," J. Math. Anal. Appl., 38(1972) 533-552.

[8] Luenberger, D.G., Introduction to Linear and Nonlinear Programming, Addison-Wesley, Reading, Mass., 1973.

[9] Ortega, J.M., and Rheinboldt, W.C., Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, New York, 1970.

[10] Ostrowski, A.M., Solution of Equations and Systems of Equations, 2nd ed., Academic Press, New York, 1966.

[11] Ralston, A., "On Differentiating Error Terms," American Mathematical Monthly 70(1963) 187-188.

[12] Tamir, A., "A One-Dimensional Search Based on Interpolating Polynomials Using Function Values Only," Management Science 22(1976) 576-586.

[13] Tamir, A., "Rates of Convergence of a One-Dimensional Search Based on Interpolating Polynomials," J. Opt. Theory Appl., 27(1979) 187-203.

[14] Traub, J.F., Iterative Methods for the Solution of Equations, Prentice-Hall, Englewood Cliffs, N.J., 1964.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

14

1. REPORT NUMBER: CCS-389	2. GOVT ACCESSION NO.: AD-A095025
	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): Unconstrained Minimization by Interpolation: Rates of Convergence.	5. TYPE OF REPORT & PERIOD COVERED: Research Report.
	6. PERFORMING ORG. REPORT NUMBER: CCS 389
7. AUTHOR(s): J. Barzilai	8. CONTRACT OR GRANT NUMBER(s): N00014-80-C-0242
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Center for Cybernetic Studies, UT Austin, Austin, TX 78712	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research (Code 434), Washington, DC	12. REPORT DATE: Dec[redacted]80
	13. NUMBER OF PAGES: 12
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): 12 16	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

This document has been approved for public release and sale; its distribution is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Unconstrained minimization, Nonpolynomial interpolation, Convergence rates

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Newton's method for finding a stationary point of $f:R^n \to R$ consists of the iteration $x_{i+1} = x_i - [\nabla^2 f(x_i)]^{-1} \cdot \nabla f(x_i)$. Its main attraction is its quadratic convergence. However, it necessitates computation and inversion of the second order derivative matrix.

406197

20. Common minimization algorithms approximate the Hessian or its inverse by first order (i.e., gradient) information. First order information algorithms in common use are known to have at best superlinear rate of convergence.

We analyze the rate of convergence of a new class of algorithms based on n-dimensional interpolation. In particular, we present a class of algorithms which use first order information only, while maintaining quadratic convergence.